PROPERTY PRICE

CASE STUDY

PROPERTY PRICE PREDICTION AND VALUATION MODEL USING MACHINE LEARNING

HIMA T. SUSEELAN
himathaivalappil@gmail.com

1. PROPERTY PRICE PREDICTION MODEL

ASSIGNMENT

- Explore the dataset given
- Build a model which predicts the listing price
- Evaluate the quality of your results
- Possible shortcoming & extensions of your approach

REGRESSION PROBLEM

Target variable is a real or continuous value

WHY ?

CASE STUDY

RELEVANCE OF PRICE PREDICTION IN REAL ESTATE MARKET

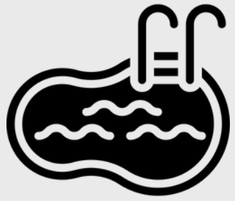
- Real estate is potentially contributing to the economic growth.
- Real estate prices are reflecting the economic level of countries, and their price ranges are of great interest to both buyers and sellers.
- Developing a price forecast model could significantly assist in predicting future property prices.

FEW FACTORS INFLUENCING PRICE OF THE PROPERTY


LOCATION


NEIGHBOURHOOD


SIZE

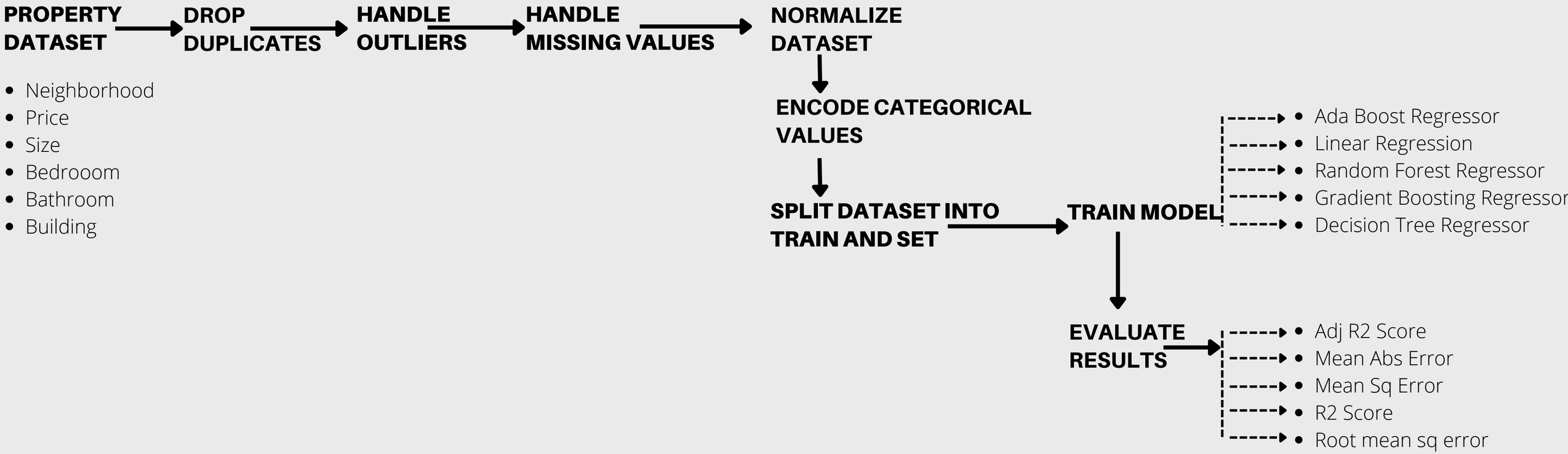

AMENITIES


SCENIC VIEW


AGE


TYPE

MIND MAPPING



INSIGHTS FROM PROPERTY DATASET

SHAPE

- 67107 Rows
- 6 Columns

- 4 numerical columns
- 2 categorical columns

DUPLICATES

- 27099 duplicate rows present

SKEWNESS

- price: 74.019004
- size: 65.173018

- Value > 1 or < -1 indicates a highly skewed distribution

MISSING VALUES

- Total 4201 missing values

- Bathrooms: 2069
- Building: 2132

SCALE

- Minimum value in dataset: -755

- Maximum value in dataset: 565352964

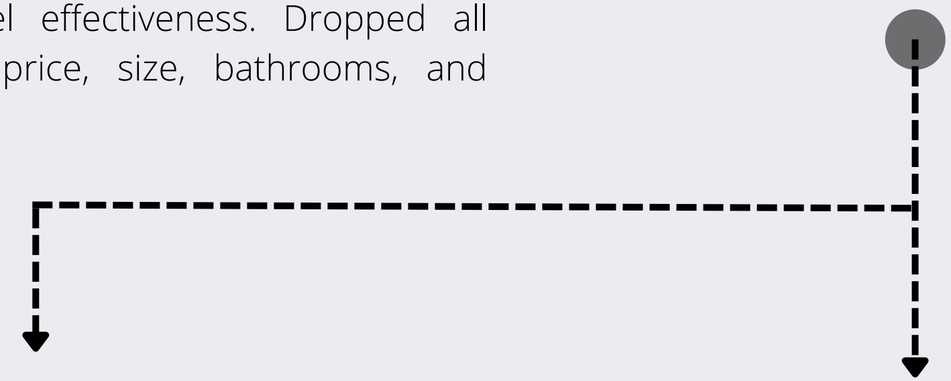
DATA PREPROCESSING

STEP 1. **DUPLICATES** → STEP 2. **OUTLIERS** → STEP 3. **MISSING DATA** →

Duplicate entries can ruin the split between train, validation and test sets. Dropped 27099 duplicate rows present in the dataset.

Outliers represent measurement errors, data entry or processing errors, or poor sampling. Outliers can skew results, and anomalies in training data can impact overall model effectiveness. Dropped all outliers present in price, size, bathrooms, and bedrooms.

Missing data can reduce the statistical power of a study and can produce biased estimates, leading to invalid conclusions.



BATHROOMS

No. of bathrooms will be always related to no. of bedrooms. Found that 71.33% of properties have (bedroom+1) no. of bathrooms. Replaced all null values with this approach.

BUILDINGS

Striped string 'Building_' and converted to numeric value. Filled missing values with most frequent value in the column.

DATA PREPROCESSING

STEP 4. **SCALING**

Scaled all values to from 0 to 1

STEP 5. **ENCODING**

One hot encoded column 'neighbourhood' into integers.

STEP 6. **MEMORY OPTIMIZATION**

Downcasted datatypes not required

MODEL TRAINING

STEP 7. **INPUT**

- size
- bedrooms
- bathrooms
- building
- PLY
- ZMS
- SNR

TARGET

- price

STEP 8. **DATA SPLIT**

Split dataset into train and test with ratio 75:25

Train data size: (18957, 8)
Test data size: (6319, 8)

STEP 9. **MODELS**

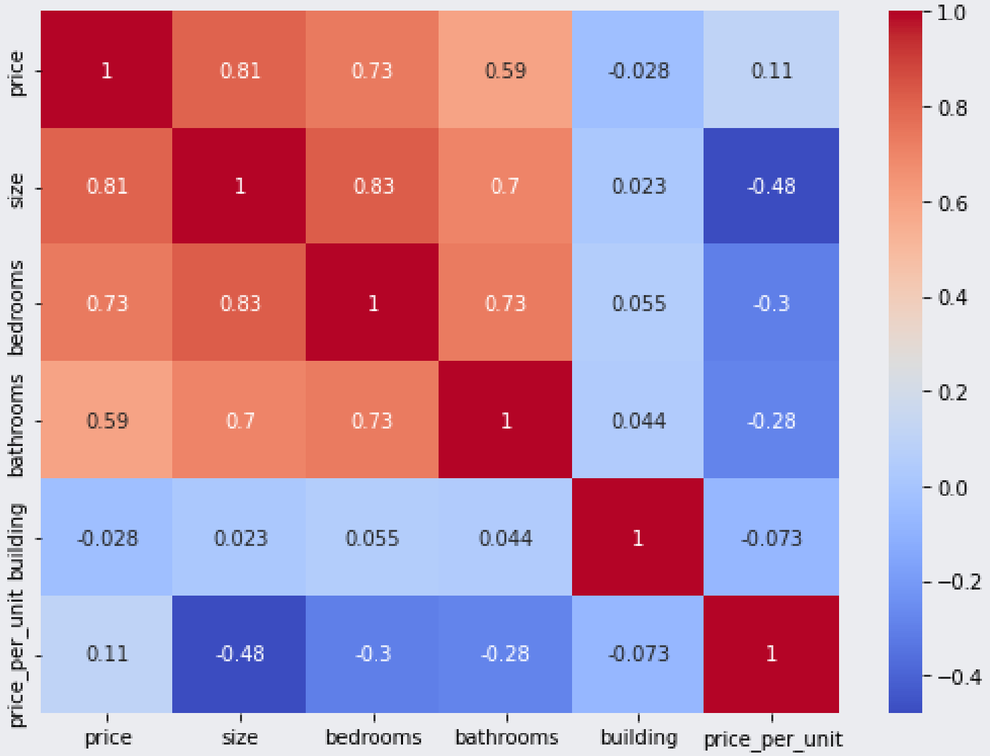
- Linear Regression
- AdaBoostRegressor
- GradientBoostingRegressor
- RandomForestRegressor
- DecisionTreeRegressor

PREDICTED PRICE

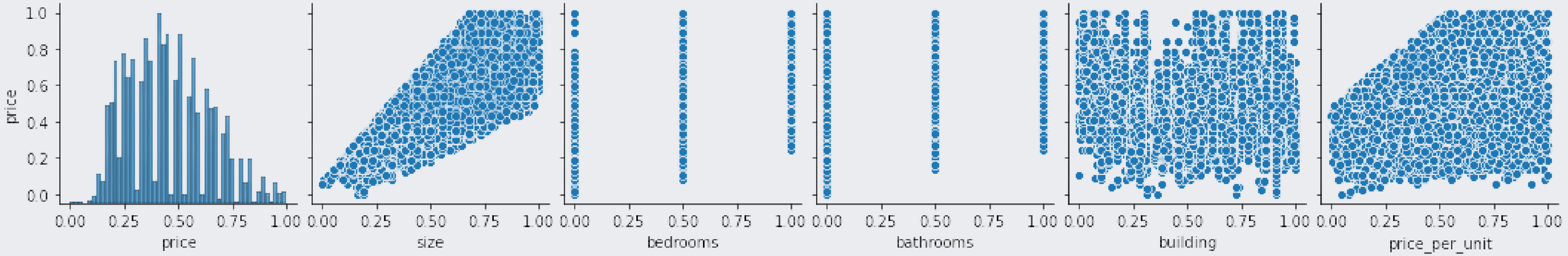
RELATIONSHIP BETWEEN COLUMNS

PAIRWISE CORRELATION OF COLUMNS PRESENTED AS A HEATMAP.

- Positive correlation implies that as one variable increases as the other increases as well.
- Inversely, a negative correlation implies that as one variable increases, the other decreases.
- **Shows moderate positive correlation between bathrooms, bedrooms, size and price.**



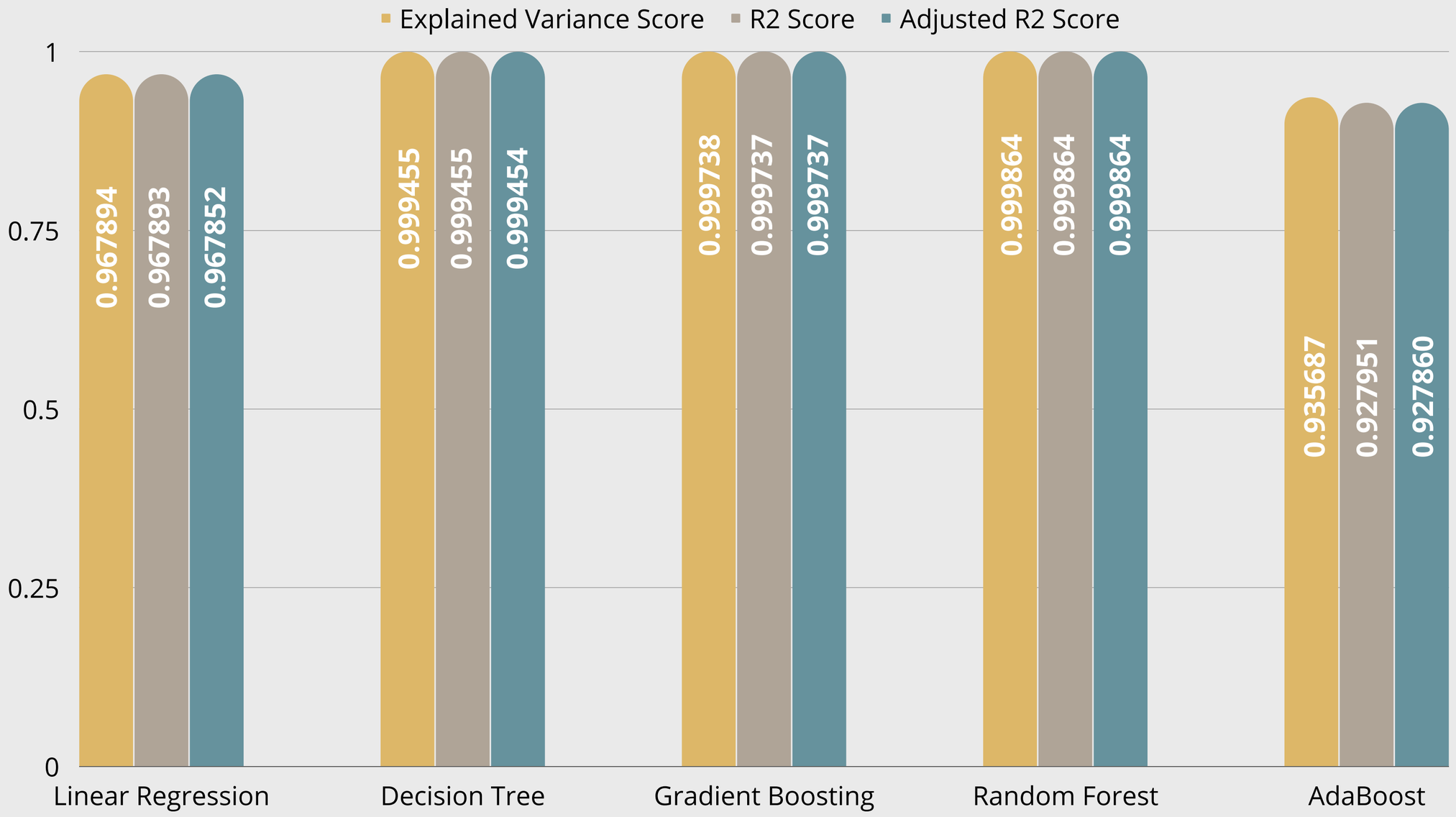
PLOTTING PAIRWISE RELATIONSHIPS W.R.T PRICE IN DATASET



PRICE PREDICTION RESULTS

From the analysis,

Random Forest Regressor works best



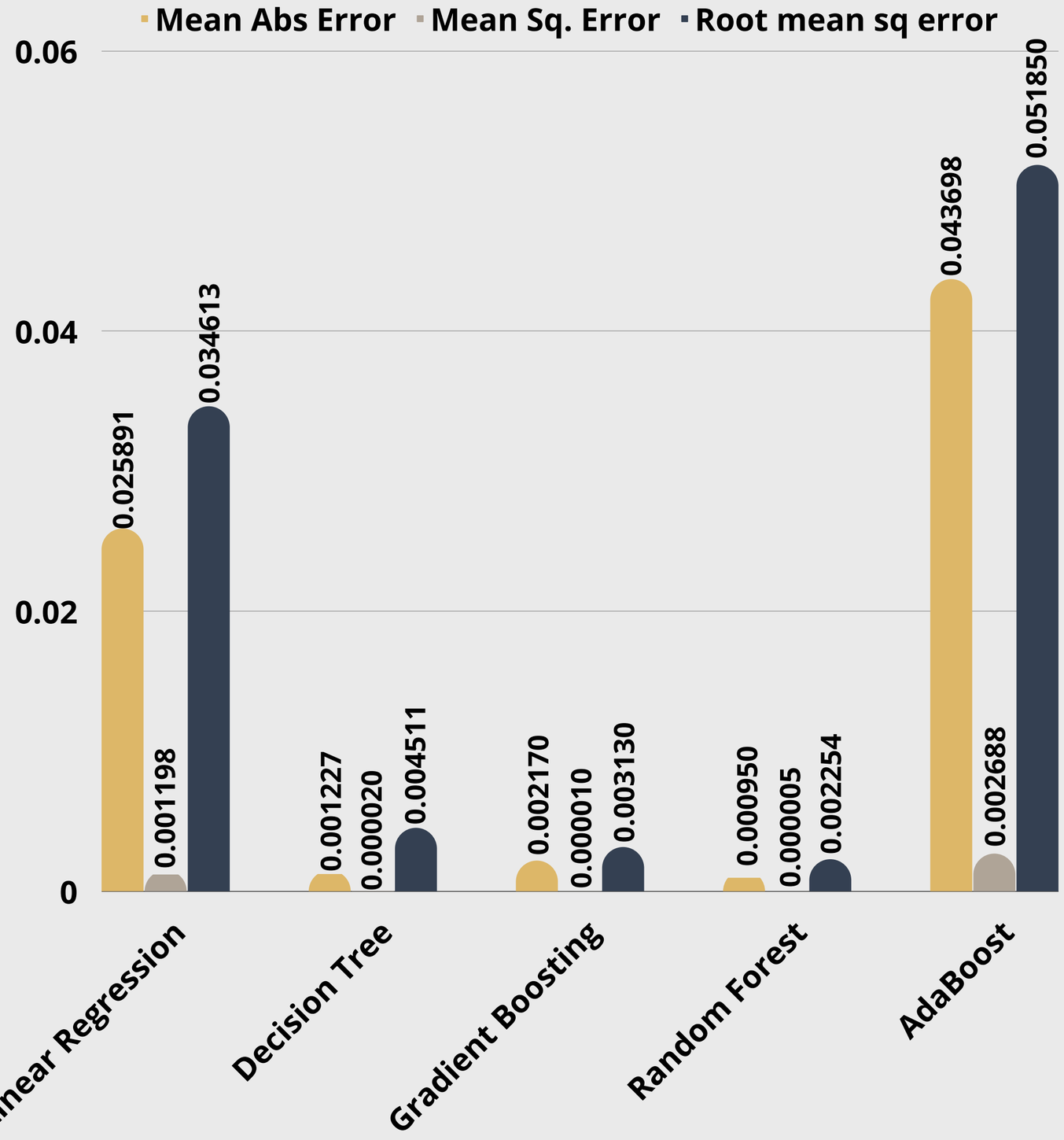
PRICE PREDICTION ERROR

- Graph shows Mean Absolute Error, Mean Square Error and Root Mean Square Error calculated for models

From the analysis,

Random Forest Regressor

works best



POSSIBLE SHORTCOMINGS

- MAY NOT BE BEST SUITED FOR OTHER RANGES OF DATA.
- SO MANY FEATURES ARE MISSING IN GIVEN DATA WHICH AFFECTS THE PROPERT PRICE

eg. Property Type, Property Age

EXTENSIONS TO THIS APPROACH

- ADD MORE ROWS WITH DIFFERENT RANGES OF DATA AND RETRAIN THE MODEL.
- ADD MORE FEATURES FOR MORE ACCURATE PREDICTIONS.
- COLLECTING HISTORICAL SALES DATA OF EACH PROPERTY, AND TRAINING WITH THOSE DATA, WE CAN PREDICT FUTURE SALE PRICES OF PROPERTIES.

2. PROPERTY PRICE VALUATION TOOL

ASSIGNMENT

- Implement a program to determine whether a property is underpriced, fairly priced or overpriced.
- Evaluate the quality of your results
- Possible shortcoming & extensions of your approach

FAIR MARKET VALUE

An estimate of the price that a property would sell for on the open market

- The best way to estimate FMV is to look at similar property sales in the same neighborhood.
- Often decided as a valuation per square foot,

PROPERTY VALUATION

STEP 1. **PREDICTION** →

Train model with target feature as price_per_sq_feet
Predict price_per_sq_feet of properties using trained regressor model

Index	Predicted Price
1	10000
2	40000
3	150000
4	22000
5	98000
6	75000
7	63000
8	55000

STEP 2. **COMPARE** →

Compare predicted price with actual prices

Index	Predicted Price	Actual Price
1	10000	12000
2	40000	35000
3	150000	120000
4	22000	22000
5	98000	120000
6	75000	74000
7	63000	68000
8	55000	56000

STEP 3. **DIFFERENCE** →

Create a new column 'Difference'.
Difference = Actual Price - Predicted Price

Index	Predicted Price	Actual Price	Difference
1	10000	12000	2000
2	40000	35000	-5000
3	150000	120000	-30000
4	22000	22000	0
5	98000	120000	22000
6	75000	74000	-1000
7	63000	68000	5000
8	55000	56000	1000

STEP 4. **DESCRIBE DIFFERENCE** → STEP 5. **CONDITIONS** → STEP 6. **PROPERTY VALUATION**

- Generate descriptive statistics of column 'difference using describe() method.
- 25th and 75th percentile can also be taken as a marker.

count	8.00000
mean	-750.00000
std	14320.31524
min	-30000.00000
25%	-2000.00000
50%	500.00000
75%	2750.00000
max	22000.00000

- But to distribute equally, taking 33.33 and 66.66 quantile values of difference.
- Here, it's -666.9 and 1666.2

Index	Predicted Price	Actual Price	Difference
1	10000	12000	2000
2	40000	35000	-5000
3	150000	120000	-30000
4	22000	22000	0
5	98000	120000	22000
6	75000	74000	-1000
7	63000	68000	5000
8	55000	56000	1000

Value between these ranges are considered as fairly priced, and below -666.9 is considered under priced and above 1666.2 is considered over priced.

Index	Predicted Price	Actual Price	Difference	Valuation
1	10000	12000	2000	over_priced
2	40000	35000	-5000	under_priced
3	150000	120000	-30000	under_priced
4	22000	22000	0	fairly_priced
5	98000	120000	22000	over_priced
6	75000	74000	-1000	under_priced
7	63000	68000	5000	over_priced
8	55000	56000	1000	fairly_priced

POSSIBLE SHORTCOMINGS

- LUXURY PROPERTIES ARE ALSO CONSIDERED SAME AS OTHER PROPERTIES, IN THE DATASET.
- FEATURES LIKE INTERIOR FINISHES, SCENIC VIEW FROM PROPERTY, FURNISHING, ACCESSIBILITY ETC ARE NOT CONSIDERED WHILE ESTIMATING THE FAIR VALUE.
- CLASSIFICATION WILL BE REFINED IF WE ADD MORE CAEGORIES LIKE SLIGHLY UNDER PRICED, SLIGHTLY OVER PRICED

EXTENSIONS TO THIS APPROACH

- WHILE DOCUMENTING THE PROPERTY, COMPUTER VISION CAN BE USED TO IDENTIFY INTERIOR FINISHES, FURNISHED/UNFURNISHED, SCENIC VIEW.
- OTHER FEATURES LIKE ACCESSIBILITY, NEAR BY PLACES CAN BE CONSIDERED USING MAP.

THANK YOU!